Supplementary Information for Unsupervised and Supervised Methodologies for Identification of Sample Pixels in Fourier Transform InfraRed Microspectroscopic Images

Xiangyu Zhao [1], Yudong Tian [1], Jingzhu Shao [1], and Chongzhao Wu [1 *]

1 Center for Biophotonics, Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

** Corresponding author: czwu@sjtu.edu.cn (Chongzhao Wu)*

Table of Contents

1. Clustering analysis for background detection

To evaluate the performance of the proposed second method, we adopt a clustering-based approach as the baseline for comparison.

This choice is motivated by the fact that the clustering method is widely used in unsupervised fingerprint analysis and shares conceptual similarities with the second method in grouping spectral features by spectral distribution. The clustering approach is applied to the same dataset used in the main manuscript. K-means clustering (KMC) is used in this work to group spectra into a predefined number of classes based on a distance metric. The resulting class images and centroid spectra were then analyzed to identify distinct spectral subtypes and their spatial distribution within the tissue. When the number of clusters is set to 2, the model can then be used to identify the sample and background pixels in the FTIR images.

As can be seen from the results, due to the limited number of hyperparameters that can be manually controlled in clustering, its flexibility and stability are relatively low, making it highly prone to misclassifying tissue pixels and background pixels. This further demonstrates the effectiveness of our proposed method, which is based on linear regression and spectral distribution analysis, in extracting tissue spectra.

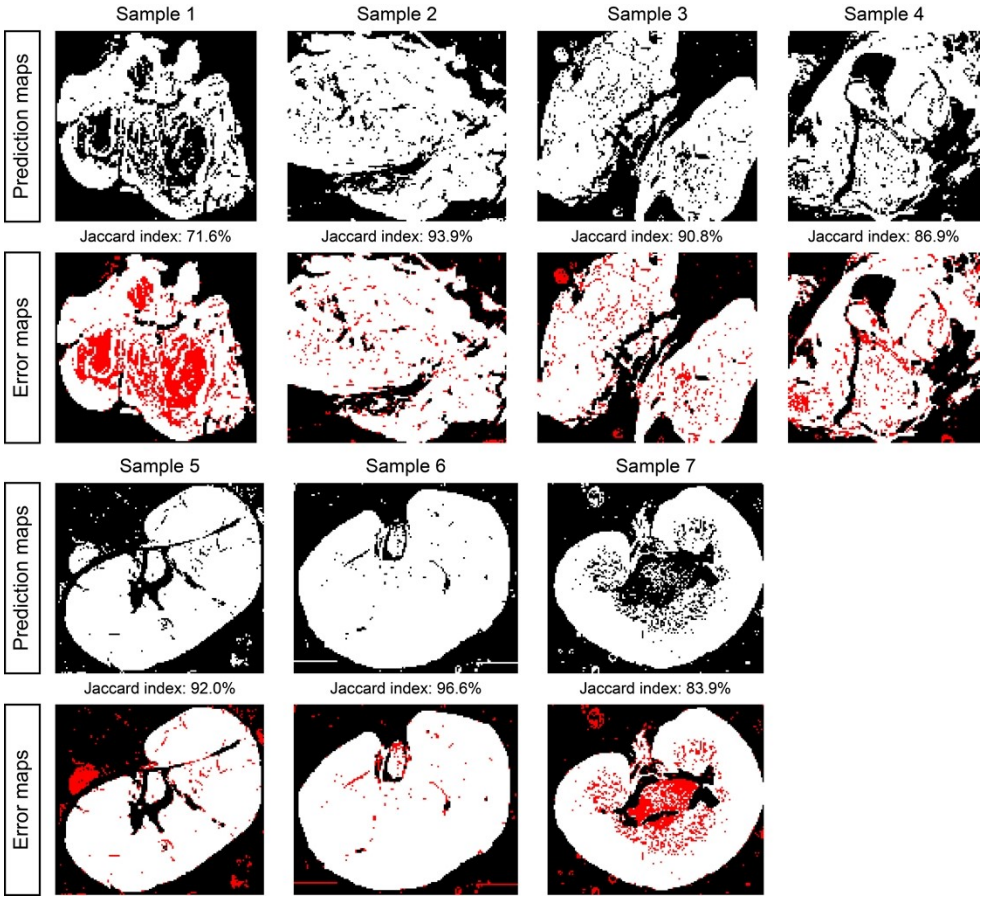

Figure S1. The results of the K-means clustering for the samples used in the manuscript.
The misclassification pixels are colored in red in the error maps.

2. Automatic paraffin detection in FTIR images proposed in previous research

In previous literature [1], an automated method based on the Extended Multiplicative Signal Correction (EMSC) [2] for paraffin pixel detection in FTIR images based on fingerprint infrared

spectra has been reported. First, EMSC was applied to IR images using the following linear model for each spectrum $A(\nu)$:

$$A(\nu) = a \cdot \bar{A}(\nu) + b \cdot P(\nu) + c \cdot I(\nu) + e(\nu) \#(S1)$$

$$b \cdot P(\nu) = b_0 + b_1\nu + b_2\nu^2 + b_3\nu^3 + b_4\nu^4 \#(S2)$$

$$c \cdot I(\nu) = \sum_{i=1}^{N_I} c_i \cdot I_i(\nu) \#(S3)$$

Here, $\bar{A}(\nu)$ is a reference spectrum chosen as the average spectrum of the FTIR image. On 1000 FTIR spectra acquired on a pure paraffin area, a principal component analysis was performed to find the main sources of spectral variability due to paraffin. The $N_I = 5$ first principal components were pooled in a matrix $I(\nu)$, named interference matrix, to model the paraffin variability into the EMSC model. $P(\nu)$ is a fourth-order Vandermonde matrix of wavenumbers used to model the baseline and light scattering effect. $e(\nu)$ is the modeling error vector. The coefficients a, $b = [b_0,b_1,b_2,b_3,b_4]$, $c = [c_1,c_2,c_3,c_4,c_5]$ are estimated by ordinary least squares. Then the natural logarithm of the modeling residue is calculated:

$$E = \ln\left(\sum e(\nu)^2\right) \#(S4)$$

Then, the automatic identification of paraffin and sample pixels were performed based on a simple application of a two cluster K-Means clustering (KMC) with parameters (a, E, b_0, c_0) to separate the pixels into two clusters, one for paraffin and the other one for tissue. It is stated that the proposed method is automatic since it does not require the setting of parameters by the operator.

Compared to the previously proposed method, our approach does not incorporate a paraffin model into the EMSC framework with the thresholding methods, addressing a broader range of background components, including contamination, paraffin, and substrate spectra, with a more stable performance. To further offer a comparison, we applied the former method to our FTIR images and visualized the results in Fig. S2.

The sample thickness for Samples 1–4 was 4 μm, and due to the stronger paraffin absorption signals in the background, the segmentation performance was generally better compared to Samples 5–7, with a thinner thickness of 2 μm. In Sample 1, the distribution of paraffin was observed in the background, likely due to artifacts introduced during the sectioning process. A similar phenomenon was also observed in Sample 5. In Sample 7, the upper-left region corresponds to a pure substrate area. When applying a two-cluster KMC approach, the model tends to classify paraffin and sample pixels into one cluster, while the pure substrate region is assigned to a separate cluster. The stability of classification for other background pixel types is compromised when threshold tuning is sacrificed to achieve automation, and the algorithm is optimized for paraffin spectral identification through the incorporation of a paraffin model. This enables them to exhibit particularly good performance in samples with very clean paraffin backgrounds (Samples 2-4, and 6).

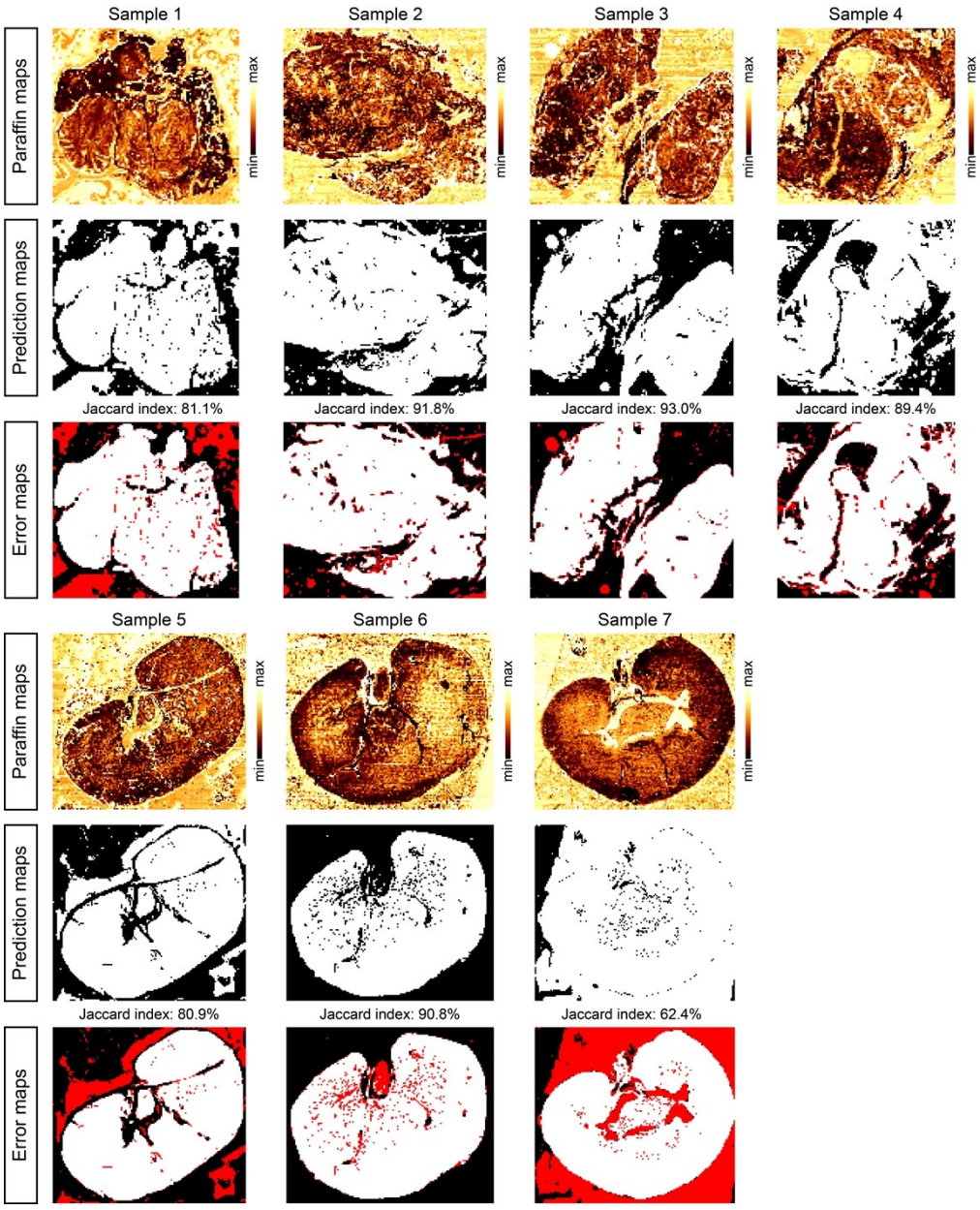

Figure S2. The results of the Extended multiplicative signal correction (EMSC)-based automatic detection of paraffin pixels. The coefficients of paraffin abundance predicted by the EMSC model are presented in the first row of the results of each sample. The misclassification pixels are colored in red in the error maps.

3. Misclassification points in the three methodologies and the comparison

To further examine the types of pixels misclassified by each method, we compare their performance in selecting sample pixel points. We highlight the misclassified pixels in red for each of the three methods.

From the comparison of the three methods applied to Sample 6, it can be observed that the integration method tends to misclassify pixels affected by baseline drift and elevation, as shown in Fig. S3(a), incorrectly assigning them to the background. In contrast, the latter two methods

do not erroneously categorize spectral features exhibiting baseline drift in the background as background spectra.

Additionally, considering that some background spectra contain features characteristic of tissue spectra, these were identified by pathologists as non-tissue regions contaminated with impurities and require careful exclusion (Samples 1,3,5, and 7). As shown in Fig. S3(b), although the Amide I/II absorption peaks, which are of interest in the integration method, are present in these spectra, the intensity of the paraffin characteristic absorption peaks (1500–1350 cm^{-1}) is significantly higher than that of the Amide I/II peaks (1700–1500 cm^{-1}), indicating that these are not genuine tissue regions. In such cases, deep learning is the most effective method for correctly classifying these background pixels.

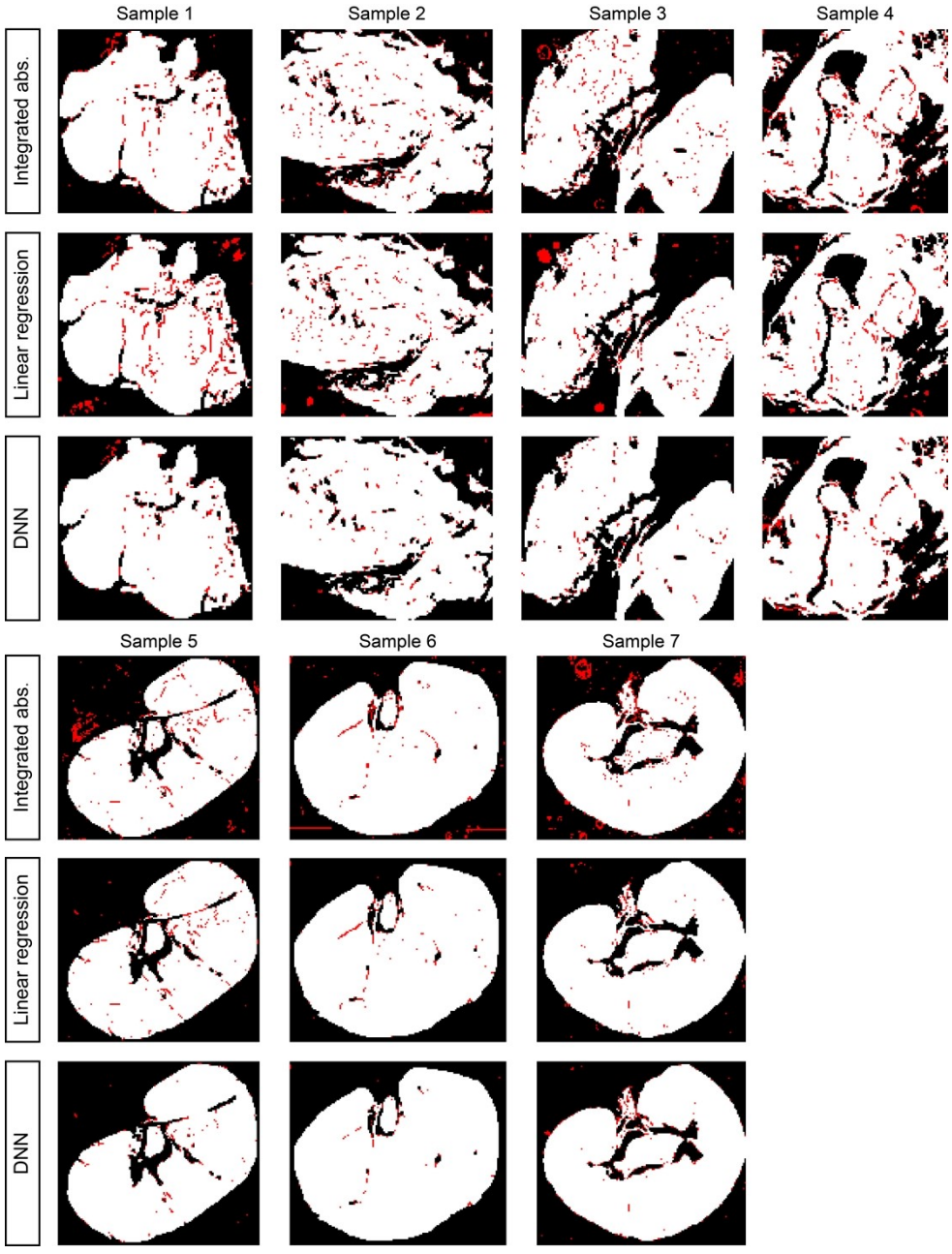

Figure S3. The error maps of the three methods on the FTIR images of all samples, indicating the misclassification points. The misclassification points are colored in red in the presented error maps.

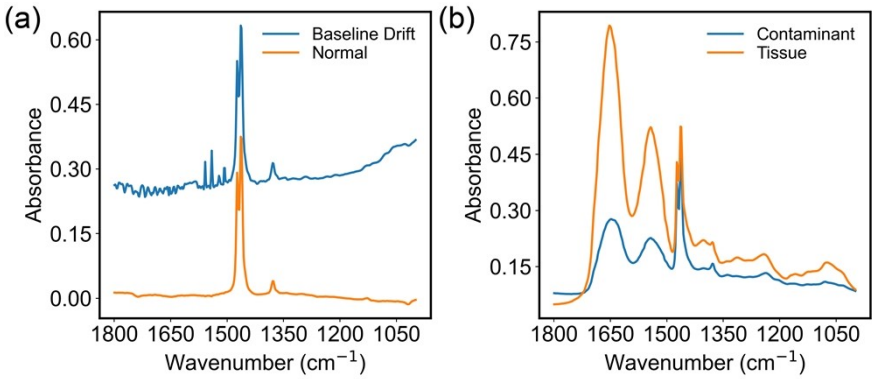

Figure S4. Spectra of the pixels that should be classified into background pixels. (a) Paraffin pixels influenced by the baseline drift extracted from Sample 6. (b) The contaminant spectrum of the non-tissue area with some similar features compared to tissue spectra.

References

(1) Boutegrabet, W.; Guenot, D.; Bouché, O.; Boulagnon-Rombi, C.; Marchal Bressenot, A.; Piot, O.; Gobinet, C. Automatic Identification of Paraffin Pixels on FTIR Images Acquired on FFPE Human Samples. *Anal. Chem.* **2021**, *93* (8), 3750–3761. https://doi.org/10.1021/acs.analchem.0c03910.

(2) Afseth, N. K.; Kohler, A. Extended Multiplicative Signal Correction in Vibrational Spectroscopy, a Tutorial. *Chemometrics and Intelligent Laboratory Systems* **2012**, *117*, 92–99. https://doi.org/10.1016/j.chemolab.2012.03.004.